PROTEIN EVOLUTION

Causes of trends in amino-acid gain and loss

Arising from: I. K. Jordan *et al.* *Nature* 433, 633–638 (2005)

Understanding how proteins evolve is important for determining the molecular basis of adaptation, for inferring phylogenies and for engineering novel proteins. It has been suggested that some amino acids were incorporated into the genetic code more recently than others[1] and, after comparing pairs of closely related genomes, Jordan *et al.*[2] report that 'recent' amino acids are becoming more common. They argue that this process has been going on since the genetic code first evolved to encompass all 20 amino acids. Here we provide evidence that the patterns observed conform with standard, nearly neutral theoretical expectations[3] and require no new explanation. This reinforces the need for caution in the interpretation of results derived from closely related taxa.

If all changes are neutral, then the equilibrium abundance of amino acids is dictated by mutation alone. At mutational equilibrium, the number of mutations causing loss of a given amino acid will equal the number of gains. But if selection skews amino-acid use, mutation will typically generate more of those amino acids that are under-represented, with this excess being selectively purged[3]. There is, however, a time lag between the mutational moderation by selection[4]: evidence for this can be found by examining the ratio of the number of synonymous substitutions per synonymous site to the number of non-synonymous substitutions per non-synonymous site, d*S*/d*N*. This ratio increases as a function of the time since common ancestry between two genomes[4] and it may explain apparently accelerated rates of short-term evolution[5–7]. Jordan *et al.*[2] notably used closely related genomes for their analysis, so might their observed trends of gain or loss of amino acids simply reflect mutation before moderation by purifying selection[3]? We provide several tests of this possibility and fail to falsify it.

First, we compared the observed profile of amino-acid gain (or loss) with that expected under mutation alone, using nucleotide changes in intergenic DNA as the mutational benchmark. We examined three bacterial groups (*Staphylococcus*, *Bacillus* and *Escherichia*/*Shigella*), considering one focal lineage in each case. We applied, through simulation, the lineage-specific mutational profile to the respective codon frequencies until the appropriate non-synonymous distance (Fig. 1, methods), and calculated the mean normalized gain–loss ratio for each amino acid. As the mutational model predicts, the simulated and observed patterns of gain and loss are similar (Fig. 1).

The mutational model also predicts that gained amino acids should occur at frequencies below mutational equilibrium. From the intergenic mutations, we approximated the equilibrium frequency of each codon as the product of estimated equilibrium levels of its nucleotides. As expected, gainers are underused (correlation between the observed normalized gain/loss ratio and mean per-codon deviation from mutational equilibrium: *Staphylococcus*: $R = -0.74$, $P = 0.00021$; *Bacillus*: $R = -0.53$, $P = 0.016$; *Escherichia*: $R = -0.50$, $P = 0.027$). As the putatively new amino acids are also those that are underused[2,8] (rank correlation between putative recruitment order[2] and per-codon deviation from equilibrium: *Staphylococcus*: $R = -0.667$, $P = 0.0013$; *Bacillus*: $R = -0.46$, $P = 0.04$; *Escherichia*: $R = -0.603$, $P = 0.0048$), mutation alone should preferentially generate the newer amino acids.

If purifying selection is moderating the mutation bias, gain–loss bias should decay

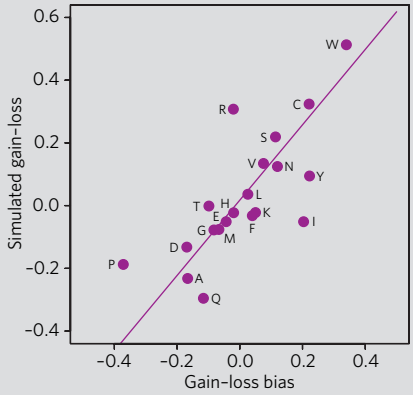

Figure 1| The observed profile of gains or losses per amino acid compared with that expected from the underlying nucleotide-level mutation bias and codon content. Data are from the focal *Staphylococcus* lineage; line indicates the principal axis regression line. Pearson product moment correlation: $R = 0.755$, $P = 0.00012$. For lineage *Bacillus*: $R = 0.502$, $P = 0.023$; for *Escherichia*, $R = 0.73$, $P = 0.0003$).
Methods. To estimate mutational profiles of the focal lineages, we compared intergenic sequence of each focal genome aligned with orthologous sequence from other genomes available for the same taxon (*Staphylococcus*, $n = 6$; *Bacillus*, $n = 5$; *Escherichia coli* / *Shigella*, $n = 5$). Only differences unique to the focal lineages were considered. The number of each of the 12 possible nucleotide changes was then divided by the frequency of the ancestral nucleotide and normalized. In the simulations, we randomly picked a codon according to its relative frequency in the focal genome, and randomly picked a base in the codon. We then mutated this base according to the mutational profile frequencies. We reiterated until the total number of non-synonymous changes reached the observed number for the focal lineage. By comparison with the initial codons, we then determined the number of gains (g) and losses (l) of each amino acid and calculated the normalized bias: $(g - l)/(g + l)$. The simulation was repeated 1,000 times.

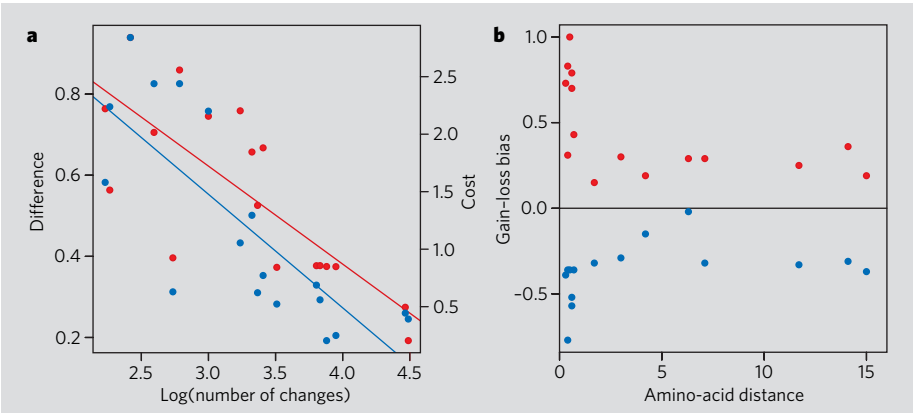

Figure 2| Relation between divergence between taxa and bias in amino-acid gain and loss. a, The difference between the normalized mean gains and losses in the strongest gainers and the strongest losers (blue) and the average cost of each amino-acid replacement (red) as a function of the log of the number of non-synonymous changes in the relevant comparisons. When the comparator species are more divergent, there is a significant decrease in the net gain or loss (blue line: $R^2 = 0.629$, $P < 0.0005$). There is a significant decrease in the average cost of each amino-acid replacement as the sequences diverge (red line: $R^2 = 0.601$, $P < 0.0005$). **b**, The gain/loss profile for proline (strong loser, blue) and cysteine (strong gainer, red) as a function of amino-acid distance in 14 diverse taxa (data from Table 1 of ref. 2). The bias is strongest in those cases in which the amino-acid distance is short and there has been little time for purifying selection. Further details are available from the authors (http://www.bath.ac.uk/bio-sci/hurst.htm).

over time[3]. Using the multiple genomes for each of our three taxa, we compared biases with the degree of divergence. We computed the average normalized gain/loss ratio for consistent gainers (S, C, F, I and V, in single-letter amino-acid notation) and consistent losers (A, D, G, P and Q), and the mean difference between these averages. As predicted, the magnitude of these biases decreases with increasing divergence over time (Fig. 2a). A diminution of bias with time is also seen in the data of Jordan *et al.* (Fig. 2b). Note that the bias need never equal zero as polymorphism will always be present.

The underuse of new amino acids may reflect their greater expense (correlation of cost[9] with putative[2] order of recruitment: R^2=0.67, P<0.0001), as biosynthetically cheaper amino acids are preferred[9] (correlation between deviation away from mutational equilibrium per codon and cost of amino acid: *Staphylococcus*: $R=-0.61$, $P=0.004$; *Bacillus*: $R=-0.43$, $P=0.058$; *Escherichia*: $R=-0.57$, $P=0.008$). Serine aside, the amino acids that are consistently gained are more expensive than those consistently lost. And there is indeed a decrease in the average cost per replacement as more diverged sequences are compared (Fig. 2a).

In sum, as expected under the nearly neutral model, mutation is biased towards the newer or costlier amino acids, but time-lagged fixation is biased towards older or cheaper amino-acid replacements. The expectation[2] of long-term bias is an artefact of extrapolation from short-term changes, and this highlights the need for caution in analyses of closely related taxa. A real bias may, however, sometimes be observed. When a lineage moves closer to the mutational equilibrium, a shift towards newer or more expensive amino acids is expected. Humans could be an example, as we may be accumulating deleterious mutations owing to reduced population size[10].

Laurence D. Hurst*, Edward J. Feil*, Eduardo P. C. Rocha†‡

*Department of Biology and Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK
e-mail: l.d.hurst@bath.ac.uk
†Atelier de BioInformatique, Université Pierre et Marie Curie, 75005 Paris, France
‡Unité GGB, Institut Pasteur, 75015 Paris, France

1. Trifonov, E. N. *J. Biomol. Struct. Dynam.* **22,** 1–11 (2004).
2. Jordan, I. K. *et al. Nature* **433,** 633–638 (2005).
3. McDonald, J. H. *Mol. Biol. Evol.* **23,** 240–244 (2006).
4. Rocha, E. P. *et al. J. Theor. Biol.* **239,** 226–235 (2006).
5. Sharp, P. M. *et al. Phil. Trans. R. Soc. Lond. B* **356,** 867–876 (2001).
6. Ho, S. Y., Phillips, M. J., Cooper, A. & Drummond, A. J. *Mol. Biol. Evol.* **22,** 1561–1568 (2005).
7. Penny, D. *Nature* **436,** 183–184 (2005).
8. Brooks, D. J., Fresco, J. R., Lesk, A. M. & Singh, M. *Mol. Biol. Evol.* **19,** 1645–1655 (2002).
9. Akashi, H. & Gojobori, T. *Proc. Natl Acad. Sci. USA* **99,** 3695–3700 (2002).
10. Keightley, P. D., Lercher, M. J. & Eyre-Walker, A. *PLoS Biol.* **3,** 282–288 (2005).

doi:10.1038/nature05137

PROTEIN EVOLUTION

Jordan et al. *reply*

Replying to: L. D. Hurst, E. J. Feil & E. P. C. Rocha *Nature* 442, doi:10.1038/nature05137 (2006)

Hurst *et al.*[1] and, earlier, McDonald[2] confirm the pattern of amino-acid gain and loss that we report[3]. However, they attribute this pattern to properties of the mutation-selection equilibrium, arguing that gainer amino acids are more common than losers among weakly deleterious, rare polymorphisms, which segregate within one or both compared species but never reach fixation. Indeed, we all[1–3] concur that gainers are, mostly, under-represented, whereas losers are over-represented with respect to mutations (Table 3 of ref. 3). Still, we cannot agree that the effect of weak negative selection is a viable alternative to our original explanation.

Hurst *et al.* show that the observed pattern is similar to that expected from the mutation process[1]. This is inconsistent with their own hypothesis in that the observed trend would be significantly weaker than the mutational bias owing to the effect of negative selection against deleterious mutations.

As we[3] and Hurst *et al.*[1] have shown, up to 30% of all substitutions contribute to the observed pattern for strong gainers and losers, even in relatively distant species at about 15% amino-acid divergence. Under the hypothesis of Hurst *et al.*, this would require about 4.5% of amino-acid sites in a genome to be occupied by deleterious alleles contributing to gain and loss. This is an order of magnitude greater than the divergence between many analysed bacterial species (Table 1 of ref. 3). Therefore, the trend cannot be explained according to Hurst *et al.*, even under the unrealistic assumptions that all deleterious polymorphism contributes to gain and loss, that all amino-acid polymorphism is deleterious, and that the divergence between bacterial species is due entirely to within-population polymorphism.

Direct comparison of amino-acid polymorphism and divergence can clarify whether the presence of non-fixed amino-acid variants is the main cause of the observed pattern. At present, comprehensive single-nucleotide-polymorphism data are available only for *Homo sapiens*. Although purifying selection is relaxed in humans compared with most other species[4,5], more than 10% of human amino-acid alleles are deleterious[5–8]. Thus, the human data can be used to analyse the effect of deleterious polymorphism on the trend. Only 2.6% of the non-synonymous differences between the public human genome and the chimpanzee genome coincide with the differences between the public genome and the Celera individual-A genome sequence. Given a similar polymorphism level in the chimpanzee, about 5% of the human–chimpanzee divergence is expected to be due to non-fixed mutations. Thus, even if all polymorphism resulted from segregation of rare deleterious gainer alleles, the observed pattern, which substantially exceeds 5% for eight out of nine strong gainers and losers, cannot be explained by polymorphism.

If mutations producing gainers and mutations eliminating losers were deleterious, levels of gain and loss in polymorphism would be substantially higher than in divergence, and the shift in allele frequency distribution would be observed in corresponding single nucleotide polymorphisms. However, the data on human single nucleotide polymorphisms do not conform to this expectation (see, for example, Table 2 of ref. 3).

We conclude that mutation-selection equilibrium is not an acceptable explanation of the universal trend in amino-acid gain and loss.

I. K. Jordan*, F. A. Kondrashov†, I. A. Adzhubei‡, Y. I. Wolf*, E. V. Koonin*, A. S. Kondrashov*, S. Sunyaev‡

*National Center for Biotechnology Information, National Institutes of Health, Bethesda, Maryland 20894, USA
†Section of Evolution and Ecology, University of California, Davis, California 95616, USA
‡Division of Genetics, Department of Medicine, Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA
e-mail: ssunyaev@rics.bwh.harvard.edu

1. Hurst, L. D., Feil, E. J. & Rocha, E. P. C. *Nature* **442,** doi:10.1038/nature05137 (2006).
2. McDonald, J. H. *Mol. Biol. Evol.* **23,** 240–244 (2006).
3. Jordan, I. K. *et al. Nature* **433,** 633–638 (2005).
4. Eyre-Walker, A., Keightley, P. D., Smith, N. G. & Gaffney, D. *Mol. Biol. Evol.* **19,** 2142–2149 (2002).
5. Chimpanzee Sequencing and Analysis Consortium *Nature* **437,** 69–87 (2005).
6. Fay, J. C., Wyckoff, G. J. & Wu, C. I. *Genetics* **158,** 1227–1234 (2001).
7. Sunyaev, S. *et al. Hum. Mol. Genet.* **10,** 591–597 (2001) .
8. Yue, P. & Moult, J. *J. Mol. Biol.* **356,** 1263–1274 (2006).

doi:10.1038/nature05138